CDA - Taller 2

Objetivo

Evaluar las capacidades del estudiante para desarrollar modelos analíticos supervisados contemplando las etapas de la metodología ASUM-DM, como son el análisis y entendimiento de los datos, preparación de datos, creación de modelos, evaluación y análisis de resultados.

Planteamiento del problema

La alcaldía de una ciudad está interesada en implementar un nuevo servicio de patinetas para incentivar la movilidad sostenible y como parte de este proceso se encuentra evaluando la viabilidad económica e impacto de dicho servicio. Dado lo anterior, le ha encargado a la consultora Andes CDA el desarrollo de un modelo predictivo de la demanda de patinetas por día con base en datos de una ciudad vecina. Su objetivo como consultor de esta empresa es la creación del mejor modelo posible de regresión lineal que le permita predecir el número de usuarios promedio por día del nuevo servicio, al mismo tiempo de poder entender la importancia y relación de las variables consideradas. A continuación, se relaciona el diccionario de datos:

Field	Description
Instant	Record Index
Date	Date (Format: YYYY-MM-DD)
Season	Season of the year
Holiday	Is it holiday?
Weather	Description of the weather situation
Temp	Temperature in Celsius
Feel_Temp	Feeling temperature in Celsius
Hum	Normalized humidity
Wind	Wind speed in m/s
Casual	Count of casual users
Registered	Count of registered users
cnt	Count of total rental bikes including both casual and registered

Actividades

A continuación, se describen más a fondo los hitos mínimos esperados por Andes CDA

Limpieza y preparación de datos (25 pts)

Búsqueda y corrección de valores atípicos, valores faltantes y duplicados; ya que el dataset no es muy extenso deberá abstenerse de eliminar registros. Así mismo, genere nuevas variables con base en la información suministrada.

Análisis de datos (15 pts)

Analice las variables del dataset, entienda sus distribuciones y correlaciones, utilice ayudas visuales.

Desarrollo de modelos de Machine Learning (30 pts)

Implemente al menos 3 modelos basados en el algoritmo de regresión lineal: uno simple, uno polinomial y uno con algún tipo de regularización.

Evaluación de modelos (20 pts)

Con base en el desempeño de cada uno de los modelos anteriormente desarrollados, concluya cuál es el modelo que se le debe presentar al cliente y exponga sus razones.

Interpretación (10 pts)

El día de la presentación del modelo ante el cliente una persona de la alcaldía le hace las siguientes preguntas:

- ¿Cuáles son las 3 variables más importantes para la predicción de la cantidad de usuarios?
- Describa cual es el escenario ideal para el incremento de usuarios
- ¿Qué pasos adicionales deberían tener en cuenta para una próxima iteración/mejora del modelo?

Ayuda

- Al consultar a un meteorólogo el cual nos sugiere agrupar las precipitaciones en un grupo, la neblina en otro grupo y el resto en otro.
- Es de interés particular el comportamiento de los usuarios durante la semana.
- Realice imputación y corrección de variables según el sentido común y la lógica del dataset.

Criterios de aceptación

- El taller debe ser desarrollado individualmente.
- Debe ser entregado en los tiempos estipulados y solo a través de BloqueNeón. No se admiten entregas por otros medios como correo electrónico.

- El entregable debe consistir en un notebook subido a un repositorio público de GitHub, el cual debe incluir los outputs de la ejecución de cada celda, pero también deberá poder ser ejecutado en su totalidad. En BloqueNeón se debe subir solo la URL del repositorio, no se admitirán commits posteriores a la fecha máxima de entrega.
- Dentro del notebook, haga uso de celdas de texto tipo markdown para exponer sus resultados y/o conclusiones de cada punto. También puede utilizar el archivo Readme del repositorio para concluir lo que considere necesario.
- Debe utilizar únicamente el dataset provisto en este taller.